Split-Plot and Multi-Stratum Designs for Statistical Inference

Luzia A. Trinca*
Department of Biostatistics, São Paulo State University, Botucatu, Brazil
and
Steven G. Gilmour
Department of Mathematics, King's College London, UK

March 16, 2017

Abstract

It is increasingly recognized that many industrial and engineering experiments use split-plot or other multi-stratum structures. Much recent work has concentrated on finding optimum, or near-optimum, designs for estimating the fixed effects parameters in multi-stratum designs. However, often inference, such as hypothesis testing or interval estimation, will also be required and for inference to be unbiased in the presence of model uncertainty requires pure error estimates of the variance components. Most optimal designs provide few, if any, pure error degrees of freedom. Gilmour and Trinca (2012) introduced design optimality criteria for inference in the context of completely randomized and block designs. Here these criteria are used stratum-by-stratum in order to obtain multi-stratum designs. It is shown that these designs have better properties for performing inference than standard optimum designs. Compound criteria, which combine the inference criteria with traditional point estimation criteria, are also used and the designs obtained are shown to compromise between point estimation and inference. Designs are obtained for two real split-plot experiments and an illustrative split-split-plot structure.

Keywords: A-optimality; D-optimality; hard-to-change factor; hard-to-set factor; mixed model; response surface; split-split-plot design.

*The authors gratefully acknowledge financial support from UNESP grant numbers PDI/028900413/PROPG-CDC and PDI/828900413/PROPG-CDC and from FAPESP grant number 2014/01818-0.

1 Introduction

Obtaining efficient results from experiments under limited resources and practical constraints has lead to an increasing number of studies dealing with design of experiments which involve restricted randomization due to some factors having levels which are hard to set. Such randomization restrictions generate the class of multi-stratum designs (Trinca and Gilmour, 2001; Goos and Gilmour, 2012; Trinca and Gilmour, 2015) from which the simplest special case includes regular orthogonal and nonorthogonal split-plot designs. In order to take into account randomization restrictions at each level of hardness to set or blocking, known as strata, random coefficients are included in the statistical model leading to a linear mixed model (Letsinger *et al.*, 1996; Trinca and Gilmour, 2001; Hinkelmann and Kempthorne, 2005). The mixed model includes fixed effects for the treatment factors and random effects for the experimental units in each stratum. Treatment effects and variances of random effects (variance components) are the parameters to be estimated from the data. The standard estimation procedure for the variance components is residual maximum likelihood (REML) whose estimates are then substituted into the usual generalised least squares (GLS) solutions in order to obtain fixed effects estimates and an approximate variance-covariance matrix of these estimates.

There is a large body of work on locally optimal split-plot designs, at point prior estimates of the variance components, considering the optimization of a single statistical property. See for example, Goos (2002), Goos and Vandebroek (2003), Goos *et al.* (2006), Goos and Donev (2007), Jones and Goos (2007), Jones and Goos (2009), Jones and Nachtsheim (2009), Macharia and Goos (2010), Jones and Goos (2012), Sambo *et al.* (2014) and Nguyen and Pham (2015) for comprehensive work on D, and more recently on I, optimum designs for fixed effects of split-plot and split-split-plot designs, with designs generally being produced assuming all variance component ratios are known to be 1. For the D criterion it has been shown that such designs are usually optimal, or very close to optimal, for a reasonable range of known values of the variance components. Usually the property optimized is based on the information matrix of the fixed effects parameters and so refers directly to point estimation of these parameters, or related properties such as prediction. It has been noted however that such methods produce designs that have no or little information for estimating the variance components. In light of this, Mylona *et al.* (2014) proposed Bayesian D optimum designs for the full information matrix, including the variance components. Of course, in some experiments, point estimation is the main priority and the standard criteria remain important for these cases.

On the other hand, much data analysis involves inference such as hypothesis testing and interval estimation. This requires either a strong assumption that the assumed model is known to be correct, which is unrealistic with empirical polynomial models, or pure error estimates of the variance components. In completely randomized structures, it has been standard practice since the dawn of response surface methodology in the 1950s to ensure estimation of pure error (Box and Draper, 2007; Myers *et al.*, 2009) to carry out model checking and inference, as in widely used designs such as central composite designs, Box-Behnken designs and subset designs. The most common design optimality criteria typically produce designs which allow little estimation of pure error, unless the number of runs is considerably greater than the number of parameters in the model. In particular, for the most popular single design criterion used, the D criterion, it has been found that optimal designs include too few degrees of freedom to estimate the extra parameters of the mixed model, the variance components (Mylona *et al.*, 2014), even assuming the polynomial model is correct. This may also be the cause of too often obtaining variance component estimates in the higher stratum to be zero as discussed in Goos *et al.* (2006) and Gilmour and Goos (2009). Such D-optimal designs, however, will often be excellent for point estimation of the polynomial effects.

It has been argued that inferences based on pure error variance component estimates should be preferred due to their known desirable statistical properties (Kowalski *et al.*, 2002; Vining *et al.*, 2005; Vining and Kowalski, 2008; Gilmour and Trinca, 2012). For small to medium sized experiments the number of degrees of freedom to estimate error can often be zero as we show in some examples in Section 3. Although variance components are not the parameters of primary interest in this type of experiment, the standard errors of the fixed effect estimates depend on them. In addition, powers of hypothesis tests are poor for few error degrees of freedom. Gilmour and Trinca (2012) proposed new design criteria which allow completely randomized and randomized block designs to be built which optimize for inference rather than point estimation. In this paper we propose methods for designing multi-stratum experiments that optimize for inference. Of course, since we are trying to do more than just point estimation, the designs produced for carrying out inference will require more runs than the standard optimal designs.

It is widely recognized that designs for real experiments must have several good properties (Box and Draper, 2007) and some advances have been made with multiple objectives (Lu *et al.*, 2011; Sambo *et al.*, 2014) and composite criteria (Jones and Nachtsheim, 2011; Gilmour and Trinca, 2012; Mylona *et al.*, 2014; Silva *et al.*, 2017) in the last few years in the context of factorial and response surface designs. Designs using composite criteria are also introduced in this paper and these are likely to be useful in practice. In Section 2 we justify and describe our algorithm for constructing designs. This is applied to several published applications in Section 3 and some general recommendations are given in Section 4.

2 A stratum-by-stratum strategy for building designs for inference

Multi-stratum designs have treatment factors applied in different strata of experimental units. For example, in the split-plot design the levels of some factors, the hard to set factors (HS), are randomized to whole plot units, and levels of other factors, the easy to set (ES) factors, are randomized to subplot units. We can also have strata, defined by a restriction in the randomization, which do not have factors applied, as in the case of a blocks stratum. The nested multi-stratum structure generalizes all the special cases. Treatments are the combinations of levels of all treatment factors and, as a consequence of the restricted randomization, some effects are confounded with unit effects in the higher strata. Let $\mathbf{T}$ be the $n \times t$ full treatment indicator matrix and $\mathbf{Y}$ the $n \times 1$ random response vector. For s strata, each containing n_i units within each unit of stratum $(i-1)$ ($i = 1, \ldots, s$ and $n_0 = 1$), such that $n = \prod_{i=1}^{s} n_i$, the full treatment model can be written as

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \sum_{i=1}^{s} \mathbf{Z}_i \boldsymbol{\epsilon}_i, \qquad (1)$$

where $\boldsymbol{\mu}$ is the $t \times 1$ vector of treatment means, $\mathbf{Z}_i$ is an $n \times m_i$ indicator matrix for the units in stratum i, $m_i = \prod_{j=1}^{i} n_j$, and $\boldsymbol{\epsilon}_i$ is the $m_i \times 1$ vector of random errors in stratum i. For continuous $\mathbf{Y}$ it is assumed that $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_i^2 \mathbf{I}_{m_i})$ and all random effects are uncorrelated. The model implies that the variance matrix of $\mathbf{Y}$ is $\mathbf{V}(\mathbf{Y}) = \sigma_s^2 \mathbf{V}$, where $\mathbf{V} = \sum_{i=1}^{s} \eta_i \mathbf{Z}_i \mathbf{Z}_i'$ and $\eta_i = \sigma_i^2/\sigma_s^2$.

To aid interpretation, for continuous or two-level factors we usually prefer the approximation

$$\mathbf{T}\boldsymbol{\mu} = \mathbf{X}_p \boldsymbol{\beta}, \qquad (2)$$

where $\mathbf{X}_p$ is the $n \times p$ model matrix for some low order polynomial and $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector ($p \leq t$). Similar approximations involving interactions up to an appropriate order are used for qualitative factors and similar models, but allowing for constraints among the components, are used for experiments with mixtures. All of these cases are covered by the methodology presented here. Given the ratios of variance components, the generalized least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_p'\mathbf{V}^{-1}\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{V}^{-1}\mathbf{Y}, \tag{3}$$

with variance given by

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \sigma_s^2(\mathbf{X}_p'\mathbf{V}^{-1}\mathbf{X}_p)^{-1}. \tag{4}$$

Since in practice we are not given the ratios of variance components, the usual approach is to estimate the variance components by REML and substitute these estimates into equations (3) and (4) to get fixed effect estimates and estimates of their standard errors.

For the analysis of experimental results many authors have argued for the appropriateness of using pure error variance component estimators. Vining *et al.* (2005) and Vining and Kowalski (2008) suggested using the sample variance from replicated treatments. Gilmour *et al.* (2017) showed that more robust and less biased REML variance component estimates can be obtained from using the residuals from the full treatment fixed effects model (equation (1)). Similar recommendations have been made, for slightly different reasons, for modeling the covariance structure in linear (Fitzmaurice *et al.*, 2011, p.175) and nonlinear (Pinheiro *et al.*, 2014; Latif and Gilmour, 2015) mixed models.

In the context of completely randomized and randomized block designs, Gilmour and Trinca (2012) noted that for valid inferences error variance should be estimated from pure error. They then proposed adjustments to the usual design optimality criteria in order that the design is correctly optimized for the inferences to be done. Hereafter in our notation we use the model matrix $\mathbf{X}$ which does not include the intercept, i.e. $\mathbf{X}_p = [\mathbf{1}\ \mathbf{X}]$. For example, for $s = 1$, for a $(DP)_S$ optimum unblocked design (with the intercept treated as a nuisance parameter) Gilmour and Trinca (2012) proposed minimizing $(F_{p-1,d;1-\alpha})^{p-1}/|\mathbf{X}'\mathbf{Q}_0\mathbf{X}|$, where d is the number of pure error degrees of freedom, $F_{p-1,d;1-\alpha}$ is the $1-\alpha$ quantile of the F distribution with $p-1$ numerator and d denominator degrees of freedom and $\mathbf{Q}_0 = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$. The number of degrees of freedom d is obtained by fitting the full treatment model, i.e. equation (1). Similar adjustments were proposed for other common design criteria. For a blocked design the function to be optimized, in order to get $(DP)_S$ optimum designs (with intercept and block effects treated as nuisance parameters), is $(F_{p-1,d_B;1-\alpha})^{p-1}/|\mathbf{X}'\mathbf{Q}\mathbf{X}|$, where d_B is the number of pure error degrees of freedom from the blocked design, $\mathbf{Q} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ and $\mathbf{Z}$ is the $n \times b$ matrix whose columns are indicators for blocks. These formulae are most easily obtained from an analysis using fixed block effects, though there is no restriction to fixed blocks implied in the data analysis. With random block effects, they represent the appropriate variances in the limit as the ratio of higher to lower stratum variance components tends to infinity; this allows us to exploit these results for the split-plot and other multi-stratum cases.

Here we extend the approach of Trinca and Gilmour (2015) for constructing multi-stratum designs stratum-by-stratum to criteria which optimize for inference. An immediate advantage of the stratum-by-stratum approach is that it is general for any multi-stratum structure and it does not require prior estimates of variance components. A further advantage is its straightforward extension to criteria for inference.

The modified stratum-by-stratum (MSS) approach of Trinca and Gilmour (2015) starts the design construction with the highest stratum in which factors are applied, considering the factors in that stratum and any higher blocking stratum only. Thus, in this phase we need an optimal unblocked or blocked design. Standard exchange algorithms are used for optimising the design in this stratum. In the second stratum, the units in the first stratum are considered as blocks and the full treatment matrix is the treatment matrix arising from factors in these two strata, i.e. each combination of levels of factors applied in either of these two strata represents a unique treatment. The approximating model matrix includes all the effects to be estimated in the second stratum. This process continues following these ideas until the lowest stratum design is constructed. Exchange algorithms are used in each step, but only levels of factors applied to the stratum being designed are subject to exchanges. There are many varieties of exchange algorithm and any of them can be used with the stratum-by-stratum approach. However, a simple modified Fedorov point exchange algorithm has proved effective. Although coordinate exchange algorithms can be faster with standard criteria, the inference-based criteria pose an extra penalty on them, since the treatment set used needs to be worked out after every exchange. With point exchange algorithms, the full treatment set can be labelled in the candidate set and these labels brought into the design with each exchange. Also, since we only deal with one stratum at a time, it is unusual for there to be more than about six factors in a stratum and it is with more factors that coordinate exchange algorithms really show speed benefits.

As noted in Trinca and Gilmour (2015) and emphasised very strongly in Trinca and Gilmour (2001), the stratum-by-stratum approach is not confined to any specific criterion, but can be used with any criterion at each step. Designs can be constructed in a straightforward manner by using the criteria of Gilmour and Trinca (2012), which optimise for interval estimation (or equivalently, hypothesis testing), rather than point estimation, in each stratum. In this paper, we explore this approach.

The first thing to note is that the designs in the higher strata are chosen to have optimal numbers of pure error degrees of freedom, but these numbers of degrees of freedom are not in general retained when we combine designs from different strata. This is because the effects of factors estimated in the lower strata will usually not all be estimated orthogonally to the higher stratum block effects. Thus some lower stratum effects will also be estimable via inter-block information in the higher strata and this might reduce the available pure error degrees of freedom in the higher strata.

This is not as serious a limitation as it might at first appear. First, many multi-stratum designs are such that there are relatively few higher stratum units, e.g. few whole plots. In this case, it is unrealistic to expect to be able to perform inferences on the parameters estimated in the whole plots stratum. The most that can be expected is that we can get decent point estimates of these parameters and inferences on the parameters estimated in the lowest stratum. On the other hand, if there are enough whole plots to give realistic chances of performing inference, then by ensuring we get pure error degrees of freedom in the subplots, we will also get them in the whole plots as a by-product.

The reason this happens is as follows. A nonorthogonal split-plot design can be considered as an incomplete block design, where the whole plots are blocks and the treatments are combinations of levels of the factors. Some treatment contrasts are completely confounded with blocks and information on them is recovered from the inter-block analysis. The block sum of squares is split into treatment sum of squares and residual sum of squares, the latter representing pure error in the whole plots stratum when we consider our response surface model. The pure error degrees of freedom correspond to the number of linearly independent unbiased estimators of contrasts among block effects that can be obtained (Hinkelmann and Kempthorne, 2005, p.14-16). Clearly a treatment which appears in r different blocks provides $r-1$ whole plot pure error degrees of freedom. Different treatments which appear in different blocks will provide more degrees of freedom if the estimable block contrasts are linearly independent.

Therefore, in order to estimate whole plot pure error with a large number of degrees of freedom, we require many replicated treatments, with the replicates spread across blocks as much as possible. The $(DP)_S$, $(AP)_S$ and similar criteria used in the subplot optimal design search will ensure that we obtain subplot pure error degrees of freedom. These require large numbers of replicates of treatments, which can be either in the same block or in different blocks. However, the good estimation of parameters (the determinant or trace part of the criterion) will tend to push replicates to be in different blocks, since it is well known that efficient designs tend not to have replicates within blocks. Hence, using pure error based criteria in the lower stratum will have the effect of ensuring pure error degrees of freedom in the higher stratum. The examples in the next section illustrate this.

Two warnings are worth noting. First, when there are more than two strata with treatment factors applied to them, this feature will ensure that there are sufficient pure error degrees of freedom in the higher strata, but will not ensure that they are spread between these strata in the best way. Second, if one is particularly interested in performing inference on the whole plot factor effects and is happy with only point estimation of the subplot factor effects (admittedly a strange requirement), one might be tempted to use the pure error criterion for constructing the whole plot design and a standard point estimation criterion for constructing the subplot design. However, the arguments above show that in this case the pure error degrees of freedom are likely to be lost, as the optimal subplot design will tend to avoid having replicates. Therefore, even in this situation, we should use pure error based (or compound) criteria in each stratum.

If, in the more usual situation, one is interested in inference for the subplot factor effects, but only point estimation for the whole plot factor effects, it is not clear if it is better to use the pure error criteria or the standard criteria in the whole plots stratum. This will be explored in some of the examples in the next section.

3 Applications: improved split-plot and split-split-plot designs

In this section we find efficient designs for inferences on the fixed effects, two for split-plot structures and one for a split-split-plot structure. We got the motivation from published designs but since inference is not possible with tight experiments we increased their sizes when necessary in order to allow for pure error degrees of freedom. For comparison, we obtained designs by methods that do not insist on including degrees of freedom for pure error such as the modified stratum-by-stratum approach and D_S optimum designs using JMP. Note that JMP constructs D optimum designs that are also D_S optimum when the nuisance parameter is the intercept or when fixed block effects are nuisance parameters. The only difference is in the efficiencies and here we show the D_S efficiencies. In the remainder of this paper, all D_S optimum designs were constructed using JMP, unless otherwise stated. In order to obtain more robust designs Gilmour and Trinca (2012) also proposed a compound criterion that incorporates up to four properties focused on the type of analysis to be performed, each property being given some priority weight κ. Here we will use three properties as follows: a global F-test ($(DP)_S$ efficiency), point estimation (weighted-A_S efficiency) and a test for lack of fit for the assumed polynomial model (degree of freedom efficiency), with weights equal to κ_1, κ_3, and κ_4, respectively, following the notation of Gilmour and Trinca (2012). The compound criteria then simplify to

$$\frac{|\mathbf{X}'\mathbf{Q}_0\mathbf{X}|^{\frac{\kappa_1}{p-1}}(n-d)^{\kappa_4}}{(F_{p-1,d;1-\alpha_1})^{\kappa_1}[tr\{\mathbf{W}(\mathbf{X}'\mathbf{Q}_0\mathbf{X})^{-1}\}]^{\kappa_3}}, \tag{5}$$

where $\mathbf{W}$ is a diagonal matrix of weights for the weighted-A criterion. In this formula, the determinant in the numerator and the F quantile in the denominator come from $(DP)_S$ efficiency, the trace in the denominator comes from A_S efficiency and the $(n-d)$ in the numerator comes from degree of freedom efficiency. In the following examples, we used relative weights 1, $\frac{1}{4}$ and 1 for linear, quadratic and interaction parameters, respectively, as in Trinca and Gilmour (2015).

For blocked designs the compound criterion function is similar with appropriate changes to take account of blocking, and simplifies to

$$\frac{|\mathbf{X}'\mathbf{Q}\mathbf{X}|^{\frac{\kappa_1}{p-1}}(n-b+1-d_B)^{\kappa_4}}{(F_{p-1,d_B;1-\alpha_1})^{\kappa_1}[tr\{\mathbf{W}(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\}]^{\kappa_3}}. \tag{6}$$

In the following examples we used $(DP)_S$ and a composite criterion using weights $\kappa = (1/3, 1/3, 1/3)$ for DP_S, A_S and DF efficiency. Designs for these criteria were constructed following the stratum-by-stratum approach but in order to keep the notation simple we will refer to them just as $(DP)_S$ and CP optimum designs respectively.

In practice, we would recommend using several different weights in the compound criterion and carefully studying many properties of the designs obtained; this can often help the experimental team to clarify the objectives of the experiment. One could go further and obtain the Pareto front of all designs which are not dominated by any other design - see Lu *et al.* (2011), Sambo *et al.* (2014), Jones (2013) and Borrotti *et al.* (2017) for recent work in this area. However, this does add considerably to the computational cost of finding designs.

3.1 Example 1: Protein Extraction

The objective of this experiment as first described in Trinca and Gilmour (2001) was the extraction of protein from a mixture of two sources A and B. The factors thought to affect production were the feed position, the feed flow rate, the gas flow rate and the concentrations of A and B. The second order model was chosen as an approximation to the response function. The experimenters had about 20 days to run the experiment but realised that if the feed position was to be set for each experimental run, as in a completely randomized design, only one run per day would be possible. This characterized the feed position as a hard to set (HS) factor, but once its level was set, two runs could be performed per day. Trinca and Gilmour (2001) proposed the use of 21 days (21 whole plots), each of size two, with one factor applied in stratum 1 and four factors in stratum 2. To design the experiment they fixed the treatment set for each stratum and arranged them to units following the stratum by stratum approach. This resulted in 2 and 0 degrees of freedom for pure error in strata 1 and 2, respectively, and low D_S-efficiency compared with competing designs. Goos (2002) constructed a D-optimum design and Trinca and Gilmour (2015) constructed designs by the MSS approach but none of them allowed pure error estimation in either of the strata. Since whole plots are of size two and the full second order model has $p = 21$ parameters, the original design was too small for performing inference. Thus we have added five extra whole plots and constructed four designs: D_S optimal assuming $\eta = 1$; MSS using the D_S criterion (MSS_D); $(DP)_S$ optimal; and a compromise design. In short, for constructing the designs stratum-by-stratum we follow:

1. Fix the candidate set for X_1, e.g. $(-1,\ 0,\ 1)$;

2. Generate a random nonsingular design for X_1, using $N = 26$ units. The model includes 3 parameters. Let $\mathbf{X}_1$ be the design-model matrix for the linear and quadratic effects ($p-1 = 2$);

3. Optimise, according to the criterion function chosen, the design $\mathbf{X}_1$ in (2) by performing point exchanges. For inference criteria, treatment labels are assigned to the rows of $\mathbf{X}_1$ in order to calculate PE degrees of freedom, the implicit model being based on the full treatment effects matrix $\mathbf{T}_1$. For inference criteria we use both models to optimise the design (although for this particular example the error df from both models are the same, since we are using a three-level design), otherwise we use just $\mathbf{X}_1$;

4. The rows of the matrices $\mathbf{X}_1$ and treatment labels are duplicated in order to form the 26 whole plots each of size 2;

5. Fix the candidate set for $X_2,\ X_3,\ X_4,\ X_5$, e.g. the 3^4 factorial;

6. With $\mathbf{X}_1$ fixed, generate a random blocked nonsingular initial design for X_2, X_3, X_4, and X_5. The whole plots act as blocks and thus the number of blocks is $b = 26$. The approximating model has block effects plus second order model terms for X_2, X_3, X_4 and X_5, plus the two-factor interaction terms between these factors and X_1. The total number of parameters in this model is $b + p - 1$ with $p - 1 = 18$. Let $\mathbf{X}_2$ be the design-model matrix for these 18 parameters;

7. Optimise, according to the criterion function chosen, the design $\mathbf{X}_2$ in step 6 by performing point exchanges. For PE degrees of freedom the model has block effects plus $\mathbf{T}_2$ (the full treatment effects matrix involving X_2, X_3, X_4 and X_5) plus the interactions $\mathbf{T}_1 \times \mathbf{T}_2$.

The four designs constructed for this example are shown in Supplementary Table A and in Table 1. The $(DP)_S$ optimal designs includes five whole plots which are completely replicated twice (whole plots 2&3, 4&5, 7&8, 12&13 and 22&23) and three whole plots which have the same treatment repeated twice (whole plots 6, 16 and 17). Similarly the compound optimum design has four whole plots completely replicated (whole plots 8&9, 10&11, 14&15 and 16&17) and two whole plots with the same treatment repeated twice (whole plots 12 and 24).

Table 1: $(DP)_S$ and CP designs for Example 1 with 26 whole plots of size 2, 1 HS and 4 ES three-level factors

$(DP)_S$

WP	X_1	X_2	X_3	X_4	X_5	WP	X_1	X_2	X_3	X_4	X_5	WP	X_1	X_2	X_3	X_4	X_5
1	−	−	−	−	−	10	0	−	−	−	0	18	+	−	−	−	−
1	−	+	+	−	0	10	0	−	+	+	−	18	+	+	+	+	−
2	−	−	0	+	+	11	0	−	−	+	0	19	+	−	−	−	+
2	−	+	−	−	−	11	0	−	+	−	+	19	+	+	+	0	+
3	−	−	0	+	+	12	0	−	+	+	+	20	+	−	−	+	−
3	−	+	−	−	−	12	0	0	0	+	−	20	+	+	−	−	0
4	−	−	+	−	0	13	0	−	+	+	+	21	+	−	+	−	−
4	−	+	+	+	+	13	0	0	0	+	−	21	+	+	−	−	−
5	−	−	+	−	0	14	0	0	−	0	+	22	+	−	+	+	0
5	−	+	+	+	+	14	0	+	+	+	0	22	+	+	−	+	+
6	−	0	−	0	0	15	0	0	+	+	0	23	+	−	+	+	0
6	−	0	−	0	0	15	0	+	0	−	−	23	+	+	−	+	+
7	−	0	+	−	−	16	0	+	+	−	+	24	+	0	0	+	+
7	−	+	−	+	−	16	0	+	+	−	+	24	+	+	+	0	−
8	−	0	+	−	−	17	0	+	+	0	−	25	+	0	+	−	−
8	−	+	−	+	−	17	0	+	+	0	−	25	+	+	0	0	+
9	−	+	−	−	+							26	+	+	−	0	−
9	−	+	+	0	−							26	+	+	+	−	+

CP

WP	X_1	X_2	X_3	X_4	X_5	WP	X_1	X_2	X_3	X_4	X_5	WP	X_1	X_2	X_3	X_4	X_5
1	−	−	−	0	0	10	0	−	0	−	0	18	+	−	−	−	−
1	−	+	+	−	−	10	0	0	−	0	−	18	+	0	+	+	0
2	−	−	−	+	+	11	0	−	0	−	0	19	+	−	−	−	+
2	−	+	0	0	0	11	0	0	−	0	−	19	+	−	+	−	−
3	−	−	0	+	−	12	0	−	0	−	+	20	+	−	−	+	0
3	−	0	−	0	+	12	0	−	0	−	+	20	+	+	−	−	+
4	−	−	+	−	0	13	0	−	0	+	+	21	+	−	−	0	+
4	−	+	0	+	+	13	0	+	+	+	−	21	+	+	−	+	−
5	−	−	+	0	−	14	0	−	+	+	+	22	+	−	+	+	−
5	−	+	0	−	+	14	0	+	+	−	+	22	+	0	0	−	0
6	−	−	+	0	+	15	0	−	+	+	+	23	+	−	+	−	+
6	−	0	−	−	−	15	0	+	+	−	+	23	+	+	0	0	0
7	−	0	0	−	+	16	0	0	0	+	−	24	+	0	0	0	−
7	−	+	−	+	0	16	0	+	−	−	0	24	+	0	0	0	−
8	−	0	+	+	0	17	0	0	0	+	−	25	+	0	−	−	+
8	−	+	0	0	−	17	0	+	−	−	0	25	+	+	+	+	+
9	−	0	+	+	0							26	+	+	−	+	+
9	−	+	0	0	−							26	+	+	+	−	−

Table 2: Skeleton ANOVA of designs for Example 1 in 26 whole plots with 2 subplots, 1 HS and 4 ES three-level factors

		Designs			
Stratum	Source	D_S	MSS_D	$(DP)_S$	CP
Whole Plot	Treat:	25	24	20	21
	$X_1,\ X_1^2$	2	2	2	2
	Inter-WP Info.	23	22	18	19
	PE	0	1	5	4
	Total WP	25	25	25	25
Sub-plot	Treat:	26	26	18	20
	2^{nd} order	18	18	18	18
	Lack of Fit	8	8	0	2
	PE	0	0	8	6
Total		51	51	51	51

The skeleton ANOVAs (Goos and Gilmour, 2012) with the breakdown of degrees of freedom are shown in Table 2 for these designs. In this and subsequent tables, indented degrees of freedom are a subdivision of the preceding unindented degrees of freedom. For example, for the D_S-optimum design there are 25 treatment degrees of freedom in the whole plots stratum, which break down into 2 for the polynomial terms of X_1 and 23 for inter-whole-plot information. The pure error degrees of freedom are obtained by fitting the full treatment model (equation (1) with fixed block effects) using a standard linear models package. The subplot pure error degrees of freedom are those which correspond to the residual sum of squares, while the whole plot pure error degrees of freedom are those which correspond to the extra sum of squares for whole plots, given treatments. The treatment degrees of freedom are then given by subtraction. In the whole plots stratum, with only a single three-level factor allocated and two degrees of freedom used for estimating its polynomial effects, there is no possibility of lack of fit being estimated, so the additional treatment degrees of freedom must be for estimating effects which are also estimated in the subplot stratum. This corresponds exactly to the usual inter-block information in an incomplete block design. In the subplots stratum, the extra degrees of freedom for treatments must correspond to lack of fit, since there are no other treatment effects which can be estimated in this stratum.

Even for the larger number of units the D_S and MSS_D designs allow no or few PE degrees of freedom in either stratum. The $(DP)_S$ optimum design optimizes the criterion in the lower stratum only since PE degrees of freedom in the higher stratum can be lost when constructing the blocked design in the second stratum. However it turns out that the design has 5 and 8 degrees of freedom for PE in the two strata. All designs except the $(DP)_S$ optimal design allow the fitting of some extra higher order terms if needed, as indicated by the lack of fit degrees of freedom in the subplot stratum. Note that the additional treatment degrees of freedom in the whole plot stratum are for contrasts involving the factors applied to the sub-plot stratum and so are of little practical use. As happened with $(DP)_S$ designs in completely randomized and randomized block structures in Gilmour and Trinca (2012) there is no extra term to be estimated in the lower stratum. The efficiencies of these designs in terms of the theoretical variance matrix of the fixed effect estimators for varying known values of variance components are shown in Table **??**. The need to estimate PE imposes some cost in terms of traditional efficiencies, especially in terms of D_S, but also to a lesser extent for A_S. The cost is attenuated by the compound optimal design that shows higher efficiencies and reasonable PE degrees of freedom.

3.2 Example 2: Ceramic Pipes

This example uses as motivation the ceramic pipe strength experiment presented in Vining *et al.* (2005). There were four factors: temperatures in zones 1 and 2 of a furnace, amount of binding in the formulation and grinding speed. Temperatures were HS and the experiment used 12 whole plots of size four. An equivalent estimation three-level central composite design (CCD) was used for the experiment. Vining *et al.* (2005) also gave an equivalent estimation Box-Behnken design (BB) for the same layout. Both designs allowed 2 and 21 PE degrees of freedom for whole and sub-plot strata, respectively, indicating that there are plenty of experimental resources at least in the lower stratum. On the other hand the efficiencies shown in Table 4 indicate that these resources may not be being used efficiently. As discussed at the end of Section 2, though we will have to use one of the inference-based criteria in the lower stratum, it is not clear whether we should use this or a standard criterion in the higher stratum, so we make some comparisons. We constructed three designs, one using the $(DP)_S$ criterion in each stratum, one using D_S in the first stratum and $(DP)_S$ in the second stratum, labelled $(DP)^{\star}_S$, and a compound optimum design,

Table 3: D_S and A_S efficiencies, relative to the D_S optimum design, and stratum $(DP)_S$ efficiencies, relative to the $(DP)_S$ optimum designs, of designs for Example 1 in 26 whole plots with 2 subplots, 1 HS and 4 ES three-level factors

		Designs			
Efficiency	η	D_S	MSS_D	$(DP)_S$	CP
	1	100.00	96.38	83.06	85.60
D_S	10	100.00	99.82	79.39	85.85
	100	100.00	100.81	78.33	85.98
	1	100.00	100.31	78.37	85.79
A_S	10	100.00	112.67	90.11	99.51
	100	100.00	117.69	111.70	114.63
	1	0.00	3.01	100.00	84.42
$(DP)_S^1$	10	0.00	2.93	100.00	83.53
	100	0.00	2.90	100.00	83.35
	1	0.00	0.00	100.00	84.23
$(DP)_S^2$	10	0.00	0.00	100.00	88.85
	100	0.00	0.00	100.00	90.34

[1]: $(DP)_S$-efficiency in the first stratum.

[2]: $(DP)_S$-efficiency in the second stratum.

all shown in Table 5. The $(DP)_S$ optimum design has six whole plots each completely replicated twice but only one pair of whole plots within which a treatment is replicated (whole plots 3&4). The $(DP)_S^{\star}$ optimum design, on the other hand, only has three whole plots completely replicated (whole plots 1&2, 9&10 and 11&12), but has ten treatments replicated within whole plots (two treatments in whole plots 3, 4, 6 and 8 and one in whole plots 5 and 7). The compound optimum design has three whole plots completely replicated (whole plots 1&2, 9&10 and 11&12) and one treatment replicated within a whole plot (whole plot 7). We note that another pair of whole plots are almost replicates, but differ in one coordinate (whole plots 4&5). These differences reflect the different criteria used in each stratum in a fairly natural way. The construction steps are similar to those for Example 1 and are described in the Supplementary Material.

For comparison we show the properties of these designs, the MSS_D optimum design published in Trinca and Gilmour (2015) and the D_S optimum design from Jones and Nachtsheim (2009). As shown in Table 6, even for reasonably abundant resources optimum designs based on a single variance property lack PE degrees of freedom. With a compound criterion we get a highly efficient design with a decent number of PE degrees of freedom and ample degrees of freedom to add higher order terms to the model in case of need. In this case, using the $(DP)_S$ criterion in both strata gives more PE degrees of freedom in the whole plots stratum than using D_S in the whole plots stratum and $(DP)_S$ in the subplots stratum. The efficiencies of these designs are shown in Table 4. Although some price is paid in terms of D_S efficiency in order to ensure sufficient PE degrees of freedom, this price is quite low, with all designs constructed having D_S efficiency greater than 93% though, in one case, the A_S efficiencies are lower. The contrast with Example 1 is due to having considerably more residual degrees of freedom in this case. This allows a large number of replicates to be included without damaging the estimation efficiency too much. The compound optimum design is again successful at ensuring reasonably high efficiencies for all criteria.

3.3 Example 3: a split-split-plot design

Jones and Goos (2009) constructed a D optimum split-split-plot design for six two-level factors, 2 VHS, 1 HS and 3 ES, using 8 whole plots each with 2 sub-plots each of size 2, considering all linear and linear by linear interaction effects. The design did not allow PE degrees of freedom and one interaction term of ES factors was fully estimated in stratum 2. It is clear that for proper

Table 4: D_S and A_S-efficiencies, relative to the D_S optimum design, and stratum $(DP)_S$ efficiencies, relative to the $(DP)_S$ optimum design, for Example 2 in 12 whole plots with 4 subplots, 2 HS and 2 ES three-level factors

		Designs						
Criterion	η	CCD	BB	D_S	MSS_D	$(DP)_S$	$(DP)_S^{\star}$	CP
D_S	1	35.81	56.03	100.00	98.44	95.14	93.83	99.02
	10	30.94	48.42	100.00	98.57	95.54	93.53	99.20
	100	26.30	41.16	100.00	98.59	95.60	93.49	99.23
A_S	1	42.72	61.43	100.00	100.63	86.53	96.67	99.15
	10	64.71	64.20	100.00	102.96	85.27	102.23	100.14
	100	70.90	64.67	100.00	103.36	85.03	103.28	100.29
$(DP)_S^1$	1	22.04	18.77	77.94	55.56	100.00	54.58	78.72
	10	21.58	18.37	76.83	55.11	100.00	54.98	78.22
	100	21.53	18.33	76.69	55.06	100.00	55.05	78.16
$(DP)_S^2$	1	22.82	45.79	52.20	49.99	100.00	91.40	86.80
	10	18.06	36.25	51.86	49.76	100.00	90.34	86.49
	100	14.02	28.13	51.81	49.74	100.00	90.21	86.45

*: the criteria for designing were D_S in the first stratum and $(DP)_S$ in the second stratum.

[1]: $(DP)_S$-efficiency in the first stratum.

[2]: $(DP)_S$-efficiency in the second stratum.

Table 5: Designs for Example 2 with 12 whole plots of size 4, 2 HS and 2 ES three-level factors

	$(DP)_S$				$(DP)^{\star}_S$				CP			
WP	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	−	−	−	−	−	−	−	−	−	−	−	0
1	−	−	−	+	−	−	−	+	−	−	0	+
1	−	−	+	−	−	−	0	0	−	−	+	−
1	−	−	+	+	−	−	+	+	−	−	+	+
2	−	−	−	−	−	−	−	−	−	−	−	0
2	−	−	−	+	−	−	−	+	−	−	0	+
2	−	−	+	−	−	−	0	0	−	−	+	−
2	−	−	+	+	−	−	+	+	−	−	+	+
3	−	0	−	+	−	0	−	+	−	0	−	−
3	−	0	−	+	−	0	−	+	−	0	−	+
3	−	0	0	−	−	0	+	−	−	0	0	−
3	−	0	+	0	−	0	+	−	−	0	+	0
4	−	0	−	+	−	+	−	−	−	+	−	−
4	−	0	−	+	−	+	−	−	−	+	−	+
4	−	0	0	−	−	+	+	+	−	+	0	0
4	−	0	+	0	−	+	+	+	−	+	+	+
5	−	+	−	−	0	−	−	0	−	+	−	−
5	−	+	0	0	0	−	0	+	−	+	−	+
5	−	+	+	−	0	−	+	−	−	+	0	0
5	−	+	+	+	0	−	+	−	−	+	+	−
6	−	+	−	−	0	0	0	−	0	−	−	−
6	−	+	0	0	0	0	0	−	0	−	−	+
6	−	+	+	−	0	0	+	0	0	−	0	−
6	−	+	+	+	0	0	+	0	0	−	+	0
7	0	+	−	0	0	+	−	0	0	0	−	+
7	0	+	0	+	0	+	−	0	0	0	0	−
7	0	+	+	−	0	+	0	+	0	0	+	0
7	0	+	+	+	0	+	+	−	0	0	+	0
8	0	+	−	0	+	0	−	+	0	+	−	0
8	0	+	0	+	+	0	−	+	0	+	0	+
8	0	+	+	−	+	0	0	0	0	+	+	−
8	0	+	+	+	+	0	0	0	0	+	+	+
9	+	−	−	−	+	−	−	−	+	−	−	−
9	+	−	−	+	+	−	−	+	+	−	−	+
9	+	−	+	−	+	−	+	−	+	−	+	−
9	+	−	+	+	+	−	+	+	+	−	+	+
10	+	−	−	−	+	−	−	−	+	−	−	−
10	+	−	−	+	+	−	−	+	+	−	−	+
10	+	−	+	−	+	−	+	−	+	−	+	−
10	+	−	+	+	+	−	+	+	+	−	+	+
11	+	+	−	−	+	+	−	−	+	+	−	−
11	+	+	−	+	+	+	−	+	+	+	−	+
11	+	+	0	−	+	+	+	−	+	+	+	−
11	+	+	+	0	+	+	+	+	+	+	+	+
12	+	+	−	−	+	+	−	−	+	+	−	−
12	+	+	−	+	+	+	−	+	+	+	−	+
12	+	+	0	−	+	+	+	−	+	+	+	−
12	+	+	+	0	+	+	+	+	+	+	+	+

*: the criteria used for designing were D_S in the first stratum and $(DP)_S$ in the second stratum.

Table 6: Skeleton ANOVA of designs for Example 2 in 12 whole plots with 4 subplots, 2 HS and 2 ES three-level factors

		Design						
Stratum	Source	CCD	BB	D_S	MSS_D	$(DP)_S$	$(DP)_S^\star$	CP
Whole Plot	Treat:	9	9	7	8	5	8	7
	2^{nd} order	$5+1^\dagger$	$5+1^\dagger$	5	5	5	5	5
	Lack of Fit	3	3	2	3	0	3	2
	PE	2	2	4	3	6	3	4
	Total WP	11	11	11	11	11	11	11
Sub-plot	Treat:	15	15	31	31	17	17	24
	2^{nd} order	$9-1^\dagger$	$9-1^\dagger$	9	9	9	9	9
	Lack of Fit	7	7	22	22	8	8	15
	PE	21	21	5	5	19	19	12
Total		47	47	47	47	47	47	47

*: the criteria for designing were D_S in the first stratum and $(DP)_S$ in the second stratum.

†: information worth one degree of freedom for the model in the lower stratum comes completely from the higher stratum (a linear combination of the quadratic effects of X_3 and X_4 is completely confounded with whole plot effects).

inference we need a larger design. For illustration, we added four whole plots to the layout and again, to compare different design strategies, we constructed designs following six approaches: D_S (fixing all ratios of variance components to be 1), MSS_D, $(DP)_S$, $(DP)_S^\star$, i.e. D_S in the two higher strata and $(DP)_S$ in the lowest stratum, and two compound optimum designs, CP using the same weight pattern in all strata and $CP^\dagger$ using D_S in the first two strata and CP in the last stratum. They are shown in Supplementary Table B and Tables 7 and 8. Here we see that the main replication is of complete subplots in different whole plots, the designs in Tables 7 and 8 having 9 (subplots 2&4&6, 3&5, 7&9, 10&12, 13&15, 14&16, 19&21, 20&24, with whole plots 7&8 and 2&3 completely replicated), 8 (subplots 1&3, 2&4, 7&17, 9&11, 10&12, 13&15, 14&16, 21&23, with whole plots 1&2, 5&6 and 7&8 completely replicated), 4 (subplots 1&3, 4&6, 9&11 and 20&23) and 3 (subplots 7&9, 10&12, and 14&16) respectively. The construction steps of our method are (in all phases the candidate set is a two-level factorial):

1. Generate a random nonsingular design for X_1 and X_2, using $N = 12$ units. The model includes 4 parameters, the intercept, two linear and one linear by linear interaction effects. Let $\mathbf{X}_1$ be the design-model matrix for the latter 3 parameters;

2. Optimise, according to the criterion function chosen, the design $\mathbf{X}_1$ in step 1 by performing point exchanges. For inference criteria, treatment labels are assigned to the rows of $\mathbf{X}_1$ in order to calculate PE degrees of freedom, the implicit model being based on the full treatment effects matrix $\mathbf{T}_1$. For inference criteria we use both models (in this specific case these are identical) to optimise the design, otherwise we use just $\mathbf{X}_1$;

3. The rows of the matrices $\mathbf{X}_1$ and treatment labels are replicated twice in order to form the 12 whole plots each of size 2;

4. With $\mathbf{X}_1$ fixed, generate a nonsingular random blocked initial design for X_3. The whole plots act as blocks and thus the number of blocks is $b = 12$. The approximating model has block effects plus the linear effect of X_3, plus $X_1 \times X_3$ and $X_2 \times X_3$. The total number of parameters in this model is $b + p - 1$ with $p - 1 = 3$. Let $\mathbf{X}_2$ be the design-model matrix for these 3 parameters;

5. Optimise, according to the criterion function chosen, the design $\mathbf{X}_2$ in step 4 by performing

point exchanges. For PE degrees of freedom the model has block effects plus $\mathbf{T}_2$ (the full treatment effects matrix for X_3) plus the interactions $\mathbf{T}_1 \times \mathbf{T}_2$;

6. The rows of the matrix $[\mathbf{X}_1\ \mathbf{X}_2]$ and treatment labels are replicated twice in order to form the 24 subplots each of size 2;

7. With $\mathbf{X}_1$ and $\mathbf{X}_2$ fixed, generate a nonsingular random blocked initial design for $X_4,\ X_5$ and X_6. The number of blocks is now $b = 24$. The model includes blocks plus linear effects and two-factor interactions involving $X_4,\ X_5$ and X_6, plus two-factor interactions between the groups of factors $X_1,\ X_2,\ X_3$ and $X_4,\ X_5,\ X_6$. The total number of parameters in this model is $b + p - 1$ with $p - 1 = 15$. Let $\mathbf{X}_3$ be the design-model matrix for these 15 parameters;

8. Optimise, according to the criterion function chosen, the design $\mathbf{X}_3$ in step 7 by performing point exchanges. For PE degrees of freedom the model has block effects plus $\mathbf{T}_3$ (the full treatment effects matrix involving $X_4,\ X_5$ and X_6) plus the interactions $\mathbf{T}_1 \times \mathbf{T}_3$, $\mathbf{T}_2 \times \mathbf{T}_3$ and $\mathbf{T}_1 \times \mathbf{T}_2 \times \mathbf{T}_3$.

9. As there are replicates of treatments at whole plot and sub-plot levels it may be possible to further optimise the design found in step 8 by swapping sub-plots among whole plots with the same levels of $X_1,\ X_2$ and X_3, always using the criterion of choice. Thus a constrained interchange algorithm is applied in which the blocking system is $b = 12$ blocks of size four and the model has pure linear effects of $X_3,\ X_4,\ X_5,\ X_6$ and all two-factor interactions except $X_1 \times X_2$, that is $p - 1 = 18$.

For the degrees of freedom from the designs compared (Table 9) we observe the same pattern as shown in the previous examples, i.e. point estimation criterion designs allowing no PE degrees of freedom in strata 2 and 3. The $(DP)_S$ optimum design allows 7, 2 and 9 PE degrees of freedom in strata 1, 2 and 3, respectively, but it does not support fitting higher order terms in the lowest stratum. Using the $(DP)_S$ criterion only in the lowest stratum again gives a slightly different allocation of degrees of freedom, which might seem preferable. As in most multi-stratum designs, some terms which are estimable in a lower stratum are also estimable in higher strata, due to inter-block information. We refer to this as inter-whole-plot or inter-subplot information depending on

Table 7: $(DP)_S$ and $(DP)_S^{\star}$ designs for Example 3 with 12 whole plots of size 2 and 24 subplots of size 2, 2 VHS, 1 HS and 3 ES two-level factors

		$(DP)_S$						$(DP)_S^{\star}$					
WP	SP	X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6
1	1	−	−	−	−	−	−	−	−	−	−	−	+
1	1	−	−	−	−	+	+	−	−	−	+	+	−
1	2	−	−	+	−	−	+	−	−	+	−	−	−
1	2	−	−	+	+	+	−	−	−	+	+	+	+
2	3	−	−	−	−	+	−	−	−	−	−	−	+
2	3	−	−	−	+	+	+	−	−	−	+	+	−
2	4	−	−	+	−	−	+	−	−	+	−	−	−
2	4	−	−	+	+	+	−	−	−	+	+	+	+
3	5	−	−	−	−	+	−	−	−	−	+	+	+
3	5	−	−	−	+	+	+	−	−	−	+	+	+
3	6	−	−	+	−	−	+	−	−	+	−	+	−
3	6	−	−	+	+	+	−	−	−	+	+	−	−
4	7	−	+	−	−	−	+	−	+	−	−	+	+
4	7	−	+	−	+	+	−	−	+	−	+	−	+
4	8	−	+	+	−	+	+	−	+	+	−	+	+
4	8	−	+	+	+	−	+	−	+	+	+	+	−
5	9	−	+	−	−	−	+	−	+	−	−	−	−
5	9	−	+	−	+	+	−	−	+	−	+	+	+
5	10	−	+	+	−	−	−	−	+	+	−	−	+
5	10	−	+	+	+	+	+	−	+	+	−	+	−
6	11	−	+	−	+	−	−	−	+	−	−	−	−
6	11	−	+	−	+	+	+	−	+	−	+	+	+
6	12	−	+	+	−	−	−	−	+	+	−	−	+
6	12	−	+	+	+	+	+	−	+	+	−	+	−
7	13	+	−	−	−	−	−	+	−	−	−	−	−
7	13	+	−	−	+	−	+	+	−	−	+	+	+
7	14	+	−	+	−	+	−	+	−	+	−	−	+
7	14	+	−	+	+	+	+	+	−	+	+	+	−
8	15	+	−	−	−	−	−	+	−	−	−	−	−
8	15	+	−	−	+	−	+	+	−	−	+	+	+
8	16	+	−	+	−	+	−	+	−	+	−	−	+
8	16	+	−	+	+	+	+	+	−	+	+	+	−
9	17	+	−	−	−	−	+	+	−	−	−	+	+
9	17	+	−	−	+	+	−	+	−	−	+	−	+
9	18	+	−	+	−	+	+	+	−	+	−	−	−
9	18	+	−	+	+	−	−	+	−	+	−	−	−
10	19	+	+	−	−	−	−	+	+	−	−	−	+
10	19	+	+	−	−	+	+	+	+	−	+	+	−
10	20	+	+	+	−	−	+	+	+	+	−	+	−
10	20	+	+	+	+	−	−	+	+	+	+	−	+
11	21	+	+	−	−	−	−	+	+	−	−	+	−
11	21	+	+	−	−	+	+	+	+	−	+	−	−
11	22	+	+	+	−	+	+	+	+	+	+	−	−
11	22	+	+	+	+	+	−	+	+	+	+	+	+
12	23	+	+	−	−	+	−	+	+	−	−	+	−
12	23	+	+	−	+	−	+	+	+	−	+	−	−
12	24	+	+	+	−	−	+	+	+	+	−	−	−
12	24	+	+	+	+	−	−	+	+	+	−	+	+

*: the criteria for designing were D_S in the first two strata and $(DP)_S$ in the third stratum.

Table 8: CP and $CP^{\dagger}$ designs for Example 3 with 12 whole plots of size 2 and 24 subplots of size 2, 2 VHS, 1 HS and 3 ES two-level factors

		CP						$CP^{\dagger}$					
WP	SP	X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6
1	1	−	−	−	−	−	−	−	−	−	−	−	−
1	1	−	−	−	−	+	+	−	−	−	−	−	+
1	2	−	−	+	−	−	+	−	−	+	−	+	+
1	2	−	−	+	+	+	−	−	−	+	+	−	+
2	3	−	−	−	−	−	−	−	−	−	−	−	−
2	3	−	−	−	−	+	+	−	−	−	+	+	+
2	4	−	−	+	−	+	+	−	−	+	−	−	+
2	4	−	−	+	+	−	+	−	−	+	+	−	−
3	5	−	−	−	−	+	−	−	−	−	−	−	+
3	5	−	−	−	+	+	+	−	−	−	+	+	+
3	6	−	−	+	−	+	+	−	−	+	+	−	+
3	6	−	−	+	+	−	+	−	−	+	+	+	−
4	7	−	+	−	+	−	+	−	+	−	−	−	+
4	7	−	+	−	+	+	−	−	+	−	+	+	−
4	8	−	+	+	−	−	+	−	+	+	−	−	−
4	8	−	+	+	+	−	−	−	+	+	−	+	+
5	9	−	+	−	−	+	+	−	+	−	−	−	+
5	9	−	+	−	+	−	+	−	+	−	+	+	−
5	10	−	+	+	−	+	−	−	+	+	−	+	−
5	10	−	+	+	+	−	−	−	+	+	+	+	+
6	11	−	+	−	−	+	+	−	+	−	−	+	−
6	11	−	+	−	+	+	−	−	+	−	+	−	−
6	12	−	+	+	−	−	−	−	+	+	−	+	−
6	12	−	+	+	+	+	+	−	+	+	+	+	+
7	13	+	−	−	−	−	+	+	−	−	−	−	+
7	13	+	−	−	−	+	−	+	−	−	+	+	−
7	14	+	−	+	−	−	−	+	−	+	−	−	−
7	14	+	−	+	+	+	−	+	−	+	+	+	+
8	15	+	−	−	−	−	+	+	−	−	−	+	−
8	15	+	−	−	+	−	−	+	−	−	+	−	−
8	16	+	−	+	−	−	−	+	−	+	−	−	−
8	16	+	−	+	+	+	+	+	−	+	+	+	+
9	17	+	−	−	−	+	−	+	−	−	−	+	−
9	17	+	−	−	+	−	−	+	−	−	+	+	+
9	18	+	−	+	+	+	−	+	−	+	−	+	+
9	18	+	−	+	+	+	+	+	−	+	+	+	−
10	19	+	+	−	−	−	−	+	+	−	−	−	−
10	19	+	+	−	+	+	+	+	+	−	+	−	+
10	20	+	+	+	−	+	+	+	+	+	−	+	−
10	20	+	+	+	+	−	+	+	+	+	+	−	−
11	21	+	+	−	+	−	+	+	+	−	−	−	−
11	21	+	+	−	+	+	+	+	+	−	−	+	+
11	22	+	+	+	−	−	+	+	+	+	−	−	+
11	22	+	+	+	+	+	−	+	+	+	−	+	−
12	23	+	+	+	−	+	+	+	+	−	−	+	+
12	23	+	+	+	+	−	+	+	+	−	+	−	+
13	24	+	+	−	−	−	−	+	+	+	−	−	+
13	24	+	+	−	+	−	+	+	+	+	+	−	−

†: the criteria for designing were D_S in the first two strata and CP in the third stratum.

Table 9: Skeleton ANOVA of designs for Example 3 in 12 whole plots with 2 subplots and 2 sub-subplots, 2 VHS, 1 HS and 3 ES two-level factors

Stratum	Source	Designs					
		D_S	MSS_D	$(DP)_S$	$(DP)_S^\star$	CP	$CP^\dagger$
Whole Plot	Treat:	7	7	4	7	3	3
	$X_1, X_2, X_1 \times X_2$	3	3	3	3	3	3
	Inter-WP Info.	4	4	1	4	0	0
	PE	4	4	7	4	8	8
	Total WP	11	11	11	11	11	11
Sub-plot	Treat:	12	12	10	9	8	9
	$X_3, X_1 \times X_3, X_2 \times X_3$	3	3	3	3	3	3
	Lack of Fit	1	1	1	1	1	1
	Inter-SP Info.	8	8	6	5	4	5
	PE	0	0	2	3	4	3
	Total SP	23	23	23	23	23	23
Sub-sub-plot	Treat:	24	24	15	15	17	18
	$X_4, X_5, X_6, X_1 \times X_4, \ldots, X_5 \times X_6$	15	15	15	15	15	15
	Lack of Fit	9	9	0	0	2	3
	PE	0	0	9	9	7	6
Total		47	47	47	47	47	47

*: the criteria for designing were D_S in the first two strata and $(DP)_S$ in the third stratum.

†: the criteria for designing were D_S in the first two strata and CP in the third stratum.

which stratum it appears in. The extra treatment degrees of freedom shown for all designs in stratum 1 are inter-whole-plot information for terms that come from the factors at lower strata and only one degree of freedom in stratum 2 could be used for augmenting the model in that stratum. The candidate set design points we use do not support fitting higher order terms for X_1 and X_2, and in stratum 2 the only higher order term that could be fitted is $X_1 \times X_2 \times X_3$. The compound optimum design (CP) distributes the PE and lack of fit degrees of freedom more equally and shows good properties in terms of efficiencies (Tables 10 and 11).

Table 10: D_S- and A_S-efficiencies, relative to the D_S optimum design, of split-split-plot designs in Example 3 with 12 whole plots with 2 subplots and 2 sub-subplots, 2 VHS, 1 HS and 3 ES two-level factors

			Designs					
Criterion	η_1	η_2	D_S	MSS_D	$(DP)_S$	$(DP)_S^\star$	CP	$CP^\dagger$
D_S	1	1	100.00	99.36	87.79	88.76	91.18	92.00
	1	10	100.00	100.26	86.30	85.51	91.64	92.81
	1	100	100.00	100.53	85.43	83.87	91.70	92.97
	100	1	100.00	99.40	87.33	88.57	90.90	91.84
	100	10	100.00	100.27	85.99	85.31	91.56	92.78
	100	100	100.00	100.53	85.39	83.84	91.69	92.96
A_S	1	1	100.00	99.34	82.18	86.60	90.17	90.28
	1	10	100.00	100.33	90.55	91.54	96.41	96.63
	1	100	100.00	100.07	98.37	98.36	99.50	99.54
	100	1	100.00	99.97	99.05	99.34	99.53	99.55
	100	10	100.00	100.04	98.69	98.86	99.55	99.58
	100	100	100.00	100.04	99.15	99.15	99.75	99.77

$\star$: the criteria for designing were D_S in the first two strata and $(DP)_S$ in the third stratum.

†: the criteria for designing were D_S in the first two strata and CP in the third stratum.

Table 11: $(DP)_S$ efficiencies, relative to the $(DP)_S$ optimum design, of split-split-plot designs in Example 3

			Designs					
Stratum	η_1	η_2	D_S	MSS_D	$(DP)_S$	$(DP)^{\star}_S$	CP	$CP^{\dagger}$
1	1	1	70.37	70.00	100.00	67.23	109.58	109.58
	1	10	67.70	67.59	100.00	66.34	108.27	108.27
	1	100	66.19	66.18	100.00	66.00	107.11	107.11
	100	1	66.03	66.02	100.00	65.97	106.96	106.96
	100	10	66.06	66.05	100.00	65.97	106.99	106.99
	100	100	66.03	66.03	100.00	65.97	106.97	106.97
2	1	1	0.00	0.00	100.00	224.70	305.02	216.73
	1	10	0.00	0.00	100.00	210.37	294.91	209.55
	1	100	0.00	0.00	100.00	207.03	291.31	206.99
	100	1	0.00	0.00	100.00	225.86	305.97	217.40
	100	10	0.00	0.00	100.00	210.50	295.11	209.69
	100	100	0.00	0.00	100.00	207.03	291.31	206.99
3	1	1	0.00	0.00	100.00	101.56	90.29	80.49
	1	10	0.00	0.00	100.00	98.71	93.13	83.02
	1	100	0.00	0.00	100.00	97.46	94.56	84.30
	100	1	0.00	0.00	100.00	101.92	90.52	80.70
	100	10	0.00	0.00	100.00	98.82	93.35	83.22
	100	100	0.00	0.00	100.00	97.46	94.57	84.31

$\star$: the criteria for designing were D_S in the first two strata and $(DP)_S$ in the third stratum.

$\dagger$: the criteria for designing were D_S in the first two strata and CP in the third stratum.

4 Discussion

Restrictions to settings of factor levels often occur in practical experiments. Sometimes experimenters are unaware of the consequences for the analysis and change the randomized order of treatments initially planned in order to meet the practical requirements (Jensen and Kowalski, 2012). So it is important to offer methodology for designing good experiments taking into account the constraints that arise in practice. We have shown that the standard approach for optimal designs does not allow valid inference in the analysis, in the classical sense that no unbiased estimators of variance components are available in the case of uncertainty about the fixed effects model. Other alternatives, such as equivalent estimation designs, do not use resources efficiently. In contrast, the stratum-by-stratum approach together with a careful choice of the criterion can produce designs that are very good in practice even with fairly small run sizes. This approach does not require prior values of the variance components and can be used with point exchange algorithms because its stratum-by-stratum nature does not usually face the problem of excessively large candidate sets. This is especially important for inference criteria since there is an extra cost for calculating them after each exchange made in the design. With the point exchange algorithm we can label the treatments in the candidate set and bring these forward as the design keeps changing in the optimization procedure. We have shown results for just one type of compound criterion, but these criteria are very flexible. Other weights and other combinations of criteria can be developed and used to produce attractive designs for practice.

In most experiments it is desirable to use designs which have many good properties and those produced here using compound criteria seem very attractive. This begs the question of which weights should be used to produce compound optimum designs. In the examples presented here, equal weight was given to each part of the compound criterion and the designs produced seem quite reasonable. However, in practice, whenever time allows, we would recommend experimenters to try out various weights and study all the properties of the designs produced. Likewise, we would recommend trying different strategies in terms of which criteria are used in the higher strata, to produce more designs for detailed consideration. This can be a very useful contribution to the discussion among the experimental team about what the real priorities for the experiment are.

Supplementary Material

suppMSPE.pdf a pdf file containing tables showing the designs obtained, which are not in the main paper, and the algorithm for Example 2.

codeMSPE.rar a zipped folder containing R code for all of the examples given in the paper.

References

Borrotti, M.; Sambo, F.; Mylona, K. and Gilmour, S. G. (2017). A multi-objective coordinate-exchange two-phase local search algorithm for multi-stratum experiments. *Statistics and Computing*, **27**, 469–481.

Box, G. E. P. and Draper, N. R. (2007) *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd edition. New York: Wiley.

Fitzmaurice, G. M.; Laird, N. M. and Ware, J. H. (2011). *Applied Logitudinal Analysis*, 2nd edition. Hoboken: John Wiley & Sons.

Gilmour, S. G. and Goos, P. (2009). Analysis of data from nonorthogonal multistratum designs in industrial experiments. *Applied Statistics*, **58**, 467–484.

Gilmour, S. G. and Trinca, L. A. (2012). Optimum design of experiments for statistical inference (with discussion). *Applied Statistics*, **61**, 345–401.

Gilmour, S. G.; Goos, P. and Großmann, H. (2017). Pure Error REML for analysing data from split-plot and multi-stratum designs. *In preparation.*

Goos, P. (2002). *The Optimal Design of Blocked and Split-Plot Experiments.* New York: Springer.

Goos, P. and Donev, A. N. (2007). Tailor-made split-plot designs with mixture and process variables. *Journal of Quality Technology*, **39**, 326–339.

Goos, P. and Gilmour, S. G. (2012). A general strategy for analysing data from split-plot and multistratum experimental designs. *Technometrics*, **54**, 340–354.

Goos, P. and Vandebroek, M. (2003). D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics*, **45**, 235–245.

Goos, P.; Langhans, I. and Vandebroek, M. (2006). Practical inference from industrial split-plot designs. *Journal of Quality Technology*, **38**, 162–179.

Hinkelmann, K. and Kempthorne, O. (2005) *Design and Analysis of Experiments*, volume 2. New York: Wiley.

Jensen, W. A. and Kowalski, S. M. (2012). Response surfaces, blocking and split-plot: an industrial experiment case study. *Quality Engineering*, **24**, 531–542.

Jones, B. (2013). Comment: Enhancing the search for compromise designs. *Technometrics*, **55**, 278–280.

Jones, B. and Goos, P. (2007). A candidate-set-free algorithm for generating D-optimal split-plot designs. *Applied Statistics*, **56**, 347–364.

Jones, B. and Goos, P. (2009). D-optimal design of split-split-plot experiments. *Biometrika*, **96**, 67–82.

Jones, B. and Goos, P. (2012). I-optimal versus D-optimal split-plot response surface designs. *Journal of Quality Technology*, **44**, 85–101.

Jones, B. and Nachtsheim, C. (2009). Split-plot designs: what, why, and how. *Journal of Quality Technology*, **41**, 340–361.

Jones, B. and Nachtsheim, C. (2011). Efficient Designs With Minimal Aliasing. *Technometrics*, **53**, 62–71.

Kowalski, S. M.; Cornell, J. A. and Vining, G. G. (2002). Split-plot designs and estimation methods for mixture experiments with process variables. *Technometrics*, **44**, 72–79.

Latif, A. H. M. M. and Gilmour, S. G. (2015). Transform-both-sides nonlinear models for in vitro pharmacokinetic experiments. *Statistical Methods in Medical Research*, **24**, 306–324.

Letsinger, J. D.; Myers, R. H. and Lentner, M. (1996). Response surface methods for bi-randomization structures. *Journal of Quality Technology*, **28**, 381–397.

Lu, L.; Anderson-Cook, C. and Robinson, T. J. (2011). Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier. *Technometrics*, **54**, 353–365.

Macharia, H. and Goos, P. (2010). *D*-optimal and *D*-efficient equivalent-estimation second-order split-plot designs. *Journal of Quality Technology*, **42**, 358–372.

Myers, R. H.; Montgomery, D. C. and Anderson-Cook, C. M. (2009). *Response Surface Methodology*, 3rd edition. New York: Wiley.

Mylona; K.; Goos, P. and Jones, B. (2014). Optimal design of blocked and split-plot experiments for fixed effects and variance component estimation. *Technometrics*, **56**, 132–144.

Nguyen, N.-K. and Pham, T.-D. (2015). Searching for *D*-efficient equivalent-estimation second-order split-plot designs. *Journal of Quality Technology*, **47**, 54–65.

Pinheiro, J.; Bornkamp, B.; Glimm, E. and Bretz, F. (2014) Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, **33**, 1646–1661.

Sambo, F.; Borrotti, M. and Mylona, K. (2014). A coordinate exchange two-phase local search algorithm for the D- and I-optimal designs of split-plot experiments. *Computational Statistics & Data Analysis*, **71**, 1193–1207.

Silva, M. A.; Gilmour, S. G. and Trinca, L. A. (2017). Factorial and response surface designs robust to missing observations. doi:10.1016/j.csda.2016.05.023

Trinca, L. A. and Gilmour, S. G. (2001). Multi-stratum response surface designs. *Technometrics*, **43**, 25–33.

Trinca, L. A. and Gilmour, S. G. (2015). Improved split-plot and multi-stratum response surface designs. *Technometrics*, **57**, 145–154.

Vining, G. G. and Kowalski, S. M. (2008). Exact inference for response surface designs within a split-plot structure. *Journal of Quality Technology*, **40**, 394–406.

Vining, G. G.; Kowalski, S. M. and Montgomery, D. C. (2005). Response surface designs within a split-plot structure. *Journal of Quality Technology*, **37**, 115–128.